Niveau(x) d'enseignement: 8 Numéro EDUQ : 8670188
Centre d'intérêt: 090

Bertrand, Richard G., 1949-.

"Le chercheur en éducation face à l'analyse de ses données" / par Richard Bertrand. -- Sainte-Foy : INRS-Education, 1984.
12 f. ; 28 cm.

Code : C-119.

Vedette(s)-secondaire(s): INRS-Education

Descripteur(s): Sciences de l'éducation * Chercheur * Analyse de données * Méthode d'analyse * Universitaire (Ordre d'enseignement)

Dans cette communication présentée à l'UQAM le 16 mars 1984, l'auteur tente de faire l'exégèse de l'expression: "Le chercheur en éducation face à l'analyse de ses données" en étudiant tour à tour chacun de ses termes, puis chaque groupe de termes et enfin l'expression toute entière. Il précise également les différences qu'il y a entre "analyse des données" et "traitement des données" ainsi qu'entre "approche exploratoire" et "approche confirmatoire". Il rappelle en conclusion que le chercheur a le devoir de prendre une part active à l'analyse de ses données et qu'il ne saurait laisser cette tâche au consultant.

Type de document: r

UQ 8670188

LE CHERCHEUR EN EDUCATION FACE A L'ANALYSE DE SES DONNEES.

par Richard Bertrand

C-119 Mars 1984

INRS-Education

EDUQ 8670188

"Le chercheur en éducation face à l'analyse de ses données"

par

Richard Bertrand

C-119

INRS-Éducation
2383, chemin Ste-Foy
Ste-Foy
G1V 1T1

Mars 1984

"Le chercheur en éducation face à l'analyse de ses données"

Communication présentée
au CIRADE, UQAM
le 16 mars 1984

"Le chercheur en éducation face à l'analyse de ses données" : il ne s'agit pas simplement du titre de l'exposé mais bien de tout l'exposé qui est inclus dans cette expression ou presque ...

Au cours des prochaines minutes, je vais en effet tenter de faire l'exégèse de cette expression en étudiant tour à tour chacun de ses termes, puis chaque groupe de termes et enfin l'expression toute entière. Cette procédure est, me semble-t-il, bien en accord avec les propos que tenait M. René Zazzo lors de son exposé d'hier.

Par **chercheur**, j'entends toute personne qui d'abord se pose une question et qui ensuite choisit d'adopter une démarche scientifique pour y répondre. Je reviendrai plus loin sur l'expression "démarche scientifique".

Bon le terme **éducation** maintenant ! Si vous me le permettez, je vais déjà faire une entorse à la procédure car je ne veux pas définir ce terme formellement. Cet exercice risquerait de nous mener trop loin, trop loin en tout cas du sujet qui nous intéresse ... Je dirai tout simplement qu'il s'agit de l'ensemble des disciplines qui nous réunit

tous aujourd'hui et pour lesquelles l'un ou l'autre d'entre nous a un certain intérêt, une certaine affinité, une certaine compétence.

Le **chercheur en éducation** est donc celui qui oeuvre dans l'une ou l'autre de ces disciplines, c'est par exemple l'un d'entre nous.

Par **données** , j'entends une pièce d'information quantitative ou en tout cas quantifiable, qui est recueillie en vue de l'étude d'un phénomène. Remarquez que j'ai mis l'accent ici sur la quantité. Non pas que je trouve l'information qualitative, vaine, imprécise ou inadéquate. Au contraire ce type d'information devrait compléter et aider à l'interprétation de l'information quantitative. Il s'agit tout simplement de fixer les idées : pour les fins de cet exposé le terme "données" ne fera donc référence qu'à de l'information quantitative ou quantifiable. Cela peut vouloir dire par exemple les notes d'un groupe de trente élèves à un examen de mathématiques ou encore les résultats d'un sondage.

Je vais devoir élaborer un peu plus sur l'expression suivante : **analyse des données** . Les plus perspicaces auront remarqué que je triche un peu ici car je ne parle pas de l'analyse de ses données. Je reviendrai au possessif plus tard.

L'analyse des données est une expression qui, selon notre expérience, porte quelquefois à confusion.

C'est pourquoi nous allons

- situer l'analyse des données dans la démarche scientifique d'une recherche

- indiquer les distinctions entre l'analyse et le traitement

- distinguer l'analyse exploratoire de l'analyse confirmatoire.

Pour analyser des données, il faut en avoir obtenues d'une façon ou d'une autre... Pour ce faire on utilise souvent la démarche (dite scientifique) comme guide :

- on choisit et on formule un problème de recherche (c.à.d. le chercheur se pose une question)
- on élabore un cadre théorique
- on met sur pied des hypothèses
- on détermine le plan de la recherche
- on construit des instruments de mesure
- on recueille des données à l'aide de ces instruments
- on analyse les données
- on interprète les résultats.

Il est vrai que dépendant du type de recherche, ces étapes sont plus ou moins présentes, plus ou moins explicites.

Il ne s'agit pas de suivre aveuglément cette démarche, étape par étape, pas à pas. Au cours d'une recherche, beaucoup d'interactions peuvent survenir entre ces étapes. C'est pourquoi cette démarche doit être vue comme un cadre de référence bien souvent pratique mais qui doit rester souple, malléable, adaptable.

Il n'en reste pas moins que la plupart de ces étapes sont susceptibles d'influencer l'analyse des données.

On pourrait en effet, élaborer longuement sur des problèmes de recherche mal définis, sur des cadres théoriques inexistants, sur des hypothèses mal formulées, sur des plans de recherche imprécis, et sur des instruments de mesure non fidèles ou non valides. Nous savons tous comment ces éléments peuvent interférer au niveau de l'analyse des données. Mais tel n'est pas notre centre d'intérêt aujourd'hui.

Venons-en plutôt à préciser ce que l'on entend par **"analyse des données"** . Il s'agit :

. de l'expression visuelle ou géométrique (graphique, tabulaire,...) des données,

. de la recherche d'indices numériques qui résument le mieux possible l'allure générale des données,

. de l'étude des regroupements de données,

. de la détection des valeurs aberrantes ("outliers").
et de la transformation des données, s'il y a lieu,

. de l'élaboration d'hypothèses de recherche, imprévues initialement, à l'aide d'un échantillon qu'on explore,

. du choix des méthodes statistiques adaptées à tel ou tel plan de recherche,

. des tests d'hypothèses statistiques,

. de l'estimation de paramètres.

Cette liste n'a pas bien sûr la prétention d'être complète puisque chaque ensemble de données comporte des particularités qui peuvent orienter l'analyse de bien des façons. Cependant cette liste est assez indicative, croyons-nous, des fonctions de l'analyse des données.

Il peut être étonnant en 1984 de ne voir aucune allusion à l'informatique dans cette liste. Eh bien, il ne s'agit pas d'un oubli, c'est fait exprès ! C'est que je tiens absolument à distinguer l'analyse du traitement.

Le **traitement** des données consiste surtout à utiliser des programmes informatiques pour :

- calculer de façon rapide et itérative des formules complexes dans le but d'accélérer le travail manuel et ce, en minimisant les erreurs de calcul,

- représenter graphiquement des données complexes.

Même si ce ne semble pas très réaliste et si cela risque d'aller fortement à contre-courant aujourd'hui, j'estime que bon nombre d'analyses des données pourraient se réaliser parfaitement sans aucun traitement informatique.

Je dis à contre-courant car il est rarissime, de nos jours, de trouver un article de revue faisant état d'une analyse des données sans avoir utilisé un programme informatique. Ce qui n'est pas nécessairement péjoratif à mon point de vue. Il ne s'agit pas de retourner 40 ans

en arrière. L'ordinateur est là pour y rester, j'en conviens. Et il est même très utile et essentiel à certains moments d'une analyse.

Il m'apparaît cependant que nous avons souvent abusé de progiciels tels SPSS, BMDP, SAS, OSIRIS, IMSL ... À telle enseigne qu'il est une pratique (regrettable) qui s'est développée et qui consiste à choisir une méthode d'analyse des données à partir d'une liste de programmes informatiques accessibles dans un progiciel. Je soutiens qu'il s'agit là d'une pratique aussi inadéquate que dangereuse. Vous comprendrez pourquoi après l'exposé de ce court exemple vécu. Un chercheur voulait comparer trois méthodes d'enseignement. Il a assigné au hasard soixante élèves en trois groupes G_1 , G_2 , G_3 , où G_1 et G_2 étaient 2 groupes expérimentaux (avec traitement) et G_3 un groupe contrôle. Il était intéressé à deux comparaisons distinctes: G_1 vs G_2 , et $\frac{G_1 + G_2}{2}$ vs G_3.

Or avec les données qu'il a recueillies, s'il avait choisi une méthode d'analyse à partir de progiciels comme SPSS et BMDP, il aurait dû effectuer une analyse de la covariance suivie de comparaisons a posteriori (les seules disponibles dans ces progiciels). Il en serait venu à la conclusion que $\frac{G_1 + G_2}{2}$ n'était pas différent de G_3. Alors qu'en utilisant les comparaisons orthogonales a priori, telles que spécifiées par le problème de recherche et les hypothèses, la conclusion inverse a été trouvée.

C'est pourquoi, j'insiste pour dire que le traitement doit être au service de l'analyse et l'ordinateur au service du chercheur, non l'inverse. Les méthodes d'analyse des données doivent être choisies conformément au plan de la recherche et non en fonction d'un éventuel traitement, si complexe soit-il.

Ce serait une erreur de croire, comme le craignait Kendall en 1968, que l'analyse des données deviendra un secteur des sciences informatiques. L'analyse des données existait bien avant l'apparition de l'ordinateur souligne Tukey en 1977 (et peut-être même existera-t-elle bien après sa disparition ?)

Ce dernier auteur est d'ailleurs à l'origine de la typologie "exploratoire-confirmatoire" de plus en plus connue et appréciée en analyse des données.

. **L'approche exploratoire** consiste surtout à :

- décrire, explorer les données en utilisant, notamment, des moyens graphiques, visuels, intuitifs,
- rechercher des indices numériques qui résument le mieux possible l'allure générale des données,
- étudier les regroupements éventuels des données,
- détecter des valeurs aberrantes.

. **L'approche confirmatoire** (dite aussi inférentielle) consiste à tester des hypothèses et à estimer des paramètres...

On ne doit pas opposer ces deux approches. Car loin de s'exclure, elles se complètent. L'approche exploratoire est en effet susceptible de soulever des hypothèses qu'il sera possible de tester dans une phase confirmatoire de l'analyse.

La meilleure façon de concrétiser la distinction entre ces deux approches consiste à en faire l'analogie avec une procédure judiciaire. L'analyse exploratoire est similaire à une enquête policière, au travail du détective qui explore, recherche des indices, soulève des hypothèses. L'analyse confirmatoire est analogue à un procès où les avocats tentent de prouver l'innocence ou la culpabilité d'un suspect, en faisant les présomptions que les membres du jury sont normaux, sans biais et homogènes, c'est-à-dire qu'ils ont la même importance aux fins du procès.

Voici un tableau synthèse de ces comparaisons

Exploratoire	Confirmatoire
. Approche descriptive	. Approche inférentielle
. Logique simple	. Logique rigide
. Indication des hypothèses à tester	. Test d'hypothèse
. Utilisation d'indices statistiques résistants	. Utilisation d'indices moins résistants
. Plan de recherche souple	. Plan de recherche rigoureux
. Usage des données en mains	. Usage de données exemptes d'erreur (si possible)
. Importance de l'expression visuelle ou graphique des données	. Moindre importance de l'expression visuelle ou graphique
. Avant une phase confirmatoire	. Après une phase exploratoire
. Vision intuitive des données	. Vision précise des données
. Esprit ouvert, divergent	. Esprit fermé, convergent
. Analogie avec une enquête policière	. Analogie avec un procès
. Importance du jugement du chercheur	. Importance du jugement du consultant
. Moindre importance de l'ordinateur	. Grande importance de l'ordinateur

Ajoutons à ce tableau que comme l'analyse exploratoire à une logique plus souple, plus intuitive, des concepts plus simples, et moins de conditions d'application que l'analyse confirmatoire, elle se prête donc beaucoup mieux, à mon sens, à un premier enseignement de l'analyse statistique et ce surtout pour des étudiants non-mathématiciens.

Pour terminer l'exposé, nous considérerons dans son ensemble, l'expression "le chercheur en éducation face à l'analyse de ses données". Notez qu'on a réintégré le possessif ici. Tout d'abord une affirmation. Je soutiens que le chercheur a non seulement la possibilité mais aussi le devoir de prendre une part active à l'analyse de ses données.

Il en a la possibilité grâce à l'avènement des techniques exploratoires de Tukey. Car Tukey ne s'est pas contenté de définir une nouvelle typologie, une nouvelle philosophie d'analyse des données. Il a aussi créé de nouveaux concepts, de nouvelles méthodes simples, graphiques et intuitives.

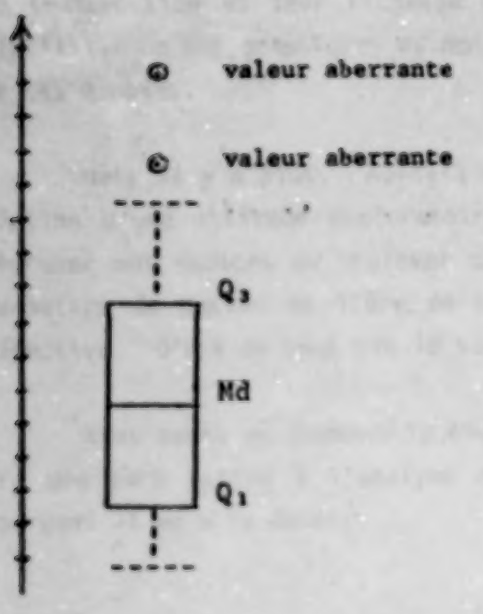
La place des graphiques en analyse exploratoire est particulièrement importante. Il s'agit pour le chercheur de scruter ses données, de les représenter visuellement avec des outils aussi simples que possible, comme par exemple le diagramme en feuilles ou le diagramme en boîte.

Le diagramme en feuilles ou "stem-and-leaf", contrairement aux graphiques classiques (histogramme, diagramme en bâtons), permet de garder les données à vue en plus de conserver les propriétés utiles des autres graphiques, comme par exemple celle de révéler l'allure générale des données.

tige		feuille	(Nb)
4	1 5 6 7 7 8 8 9		(8)
⑤	0 1 6 7 7 9		(6)
6	1 2 3 4 5 ⑥ 7		(7)
7	4 6		(2)
			(T = 23)

Le nombre situé à gauche du trait vertical de chaque ligne est une tige ou "stem". Chacune des entités numériques à droite du trait vertical est nommée feuille ou "leaf". La colonne titrée (Nb) et située à l'extrême droite du diagramme indique le nombre de feuilles pour chaque tige. Le graphe suivant comprend donc quatre tiges ayant chacune respectivement 8, 6, 7 et 2 feuilles. On compte en tout 23 feuilles qui correspondent aux 23 données, soit 41, 45, 46 ... 76.

Le diagramme en boîte ou "box-and-whiskers" complète bien le diagramme en feuilles.



Construit à l'échelle, le diagramme en boîte permet, en un seul coup d'oeil, de détecter des valeurs aberrantes et de comparer plusieurs distributions, vertus qu'on ne peut accorder au diagramme en feuilles.

Comme on peut le voir, ce type de diagramme ressemble à un rectangle, que nous appelons une boîte, ou "box", prolongé à chaque extrémité de traitillés, appelés moustaches ou "whiskers". L'axe vertical, à gauche de la boîte, indique que les valeurs représentées dans ce diagramme ont été calculées à l'échelle. Le premier (Q_1) et le troisième (Q_3) quartiles sont situés aux extrémités de la boîte alors que la position de la médiane (Md) est représentée par une barre horizontale divisant la boîte en deux parties. La dimension verticale de la boîte représente la différence réelle entre Q_3 et Q_1, soit l'étendue interquartile (EI). Toute donnée située à plus de 1,5 x EI de l'un ou l'autre des quartiles est appelée une valeur aberrante. Il s'agit d'une donnée qui se démarque suffisamment des autres pour qu'il soit nécessaire de s'y attarder de façon bien particulière.

C'est notre avis que ces diagrammes sauront, par leur simplicité de construction et leur richesse visuelle, faciliter une éclosion de l'utilisation des graphiques et donner au chercheur le goût de l'analyse de ses données.

Mais il y a plus. Au-delà des techniques et des méthodes, l'adoption d'une attitude exploratoire qui est de décrire, explorer, rechercher des indices ou soulever de nouvelles hypothèses pourrait nous permettre de passer de l'ère du chercheur-juge à l'ère du chercheur-détective. C'est en tout cas le souhait que nous émettons.

Nous avons vu comment le chercheur avait la possibilité de prendre une part active à l'analyse de ses données. Il me reste à dire pourquoi il en a le devoir.

Tout en concédant qu'un chercheur ne peut bien souvent effectuer toutes ses analyses de données de façon isolé, il m'apparaît, par ailleurs, qu'il ne doit pas les laisser à un consultant, statisticien ou informaticien, si complexes soient les analyses et si compétent soit le consultant.

Seul le chercheur connaît à fond le problème de recherche à l'étude, les hypothèses qu'il veut tester, etc... Si le chercheur donne "carte blanche" au consultant pour l'analyse de ses données, il y a un risque que l'analyse ne corresponde pas tout à fait au problème de recherche et en outre, même si elle correspondait, cette analyse ne pourra être aussi riche que si le chercheur s'implique activement.

S'il est illusoire de penser que deux ou trois petits conseils donnés par le consultant en quelques minutes peuvent suffire au chercheur pour établir son plan de recherche et faire ses analyses de données, il est tout aussi illusoire de penser qu'un consultant, étranger à la recherche, peut effectuer les analyses d'un chercheur sans implication importante de ce dernier.

À mon avis le consultant et le chercheur n'ont d'autre choix que de travailler en équipe. Ce qui peut être difficile au début, dépendant du niveau de compétence en statistique du chercheur et de la familiarité du domaine de recherche du chercheur qu'aura le consultant. Il y a encore ici un risque, mais les profits escomptés sont plus grands, beaucoup plus grands. Les médecins, les avocats ou les ingénieurs n'hésitent pas à se regrouper pour étreindre une réalité complexe. En éducation, l'établissement du plan de la recherche et l'analyse des données sont des réalités fort complexes qui exigent des efforts soutenus du maximum d'intervenants possible. Ce n'est qu'à ce prix que la recherche aura un sens.